DOCUMENT RESUME

ED 279 727 TM 870 176

AUTHOR Wise, Lauress L.; McLaughlin, Donald H.

TITLE Guidebook for Imputation of Missing Data. Technical Report No. 17.

INSTITUTION American Institutes for Research in the Behavioral Sciences. Palo Alto, CA. Statistical Analysis Group in Education.

SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Aug 80
CONTRACT 300-78-0150
NOTE 48p.; Some figures contain small print.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; Computer Software; *Data Analysis; *Data Collection; *Data Processing; Estimation (Mathematics); Mathematical Models; Regression (Statistics); *Research Methodology; Research Problems; *Statistical Analysis
IDENTIFIERS National Center for Education Statistics

ABSTRACT

This guidebook is designed for data analysts who are working with computer data files that contain records with incomplete data. It indicates choices the analyst must make and the criteria for making those choices in regard to the following questions: (1) What resources are available for performing the imputation? (2) How big is the data file? (3) What is the purpose for imputing missing data? (4) What structures exist in the recorded variables? (5) What is the pattern of missing data? (6) What assumptions are acceptable for this imputation? Answers to these questions constitute recommendations for imputation procedures. Several alternative recommendations and the conditions that determine the appropriateness of use are considered. The final section of the guidebook contains instructions for using PROC IMPUTE created by the Statistical Analysis Group in Education for the National Center for Education Statistics, and for interpreting its results. Appendices include: (1) Processing Time by Numbers of Variables and PROC IMPUTE; and (2) Sample Statistical Analysis System Program to Reweight for Total Nonresponse. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279727

Guidebook for Imputation of Missing Data

Prepared by

SAGE
STATISTICAL ANALYSIS GROUP IN EDUCATION

For the

National Center for Education Statistics

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. V. Clemans
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

American Institutes for Research

Box 1113, Palo Alto, California 94302

TM 870 176

TECHNICAL REPORT NO. 17

GUIDEBOOK FOR IMPUTATION OF MISSING DATA

Lauress L. Wise

Donald H. McLaughlin

Statistical Analysis Group in Education
American Institutes for Research
P.O. Box 1113
Palo Alto, California 94302

This guidebook was prepared by the Statistical Analysis Group in Education, American Institutes for Research, for the National Center for Education Statistics, under contract #300-78-150. Contractors undertaking such projects are encouraged to express freely their professional judgment. This guidebook, therefore, does not necessarily represent positions or policies of the Department of Education, and no official endorsement should be inferred.

August 1980

I. ALTERNATIVE IMPUTATION STRATEGIES

What This Guidebook Is About

This guidebook is for data analysts who are working with computer data files that contain records with incomplete data. Specifically, the guidebook pertains to data from surveys that had some nonresponse. The guidebook indicates choices the analyst must make and criteria for making those choices. Because dealing with missing data can be facilitated by careful design of the survey instrument and data collection, this guidebook includes useful information for survey designers as well as data analysts.

The Guidebook for Imputation was prepared as supporting documentation for a particular missing data imputation procedure developed under NCES contract, but as we shall see, the choice of best procedure depends on both the contents of the data file and the objectives of the analyst. To be more precise, the analyst must address the following six questions to decide on an imputation procedure. Each of these will be discussed in turn in this guidebook.

1. What resources are available for performing the imputation?
2. How big is the data file?
3. What is the purpose for imputing missing data?
4. What structures exist in the recorded variables?
5. What is the pattern of missing data?
6. What assumptions are acceptable for the imputation?

The answers to these questions will constitute recommendations for imputation procedures. We shall consider these in turn, and then list a series of specific alternative recommendations, indicating the conditions that determine the appropriateness of use of each of several alternative procedures. The final section of this guidebook contains instructions for using PROC IMPUTE, created by SAGE for NCES, and for interpreting its results.

Many agencies have done a substantial amount of work recently to improve imputation procedures, to which this guidebook only refers in terms of general principles and findings. Interested readers who wish to pursue alternatives other than the use of standard packages might refer to Aziz and Scheuren (1978) and Madow (1979) for compendia of different perspectives, models, procedures, and findings.

There are basically four types of imputation procedures:

1. superficial methods, such as ignoring missing data, using complete cases only, or assigning the mean or modal value for all missing cases;

2. weighting methods, in which missing values are implicitly filled in by increasing the weights assigned to similar cases that responded;

3. single-valued explicit imputation, in which a response is inserted into the data file in place of the missing data code; and

4. multi-valued explicit imputation, in which replicate files are created with different responses inserted based on different underlying imputation models.

Among the weighting methods, there are two major types, those that incorporate external information about response distributions, such as raking ratio estimators (Oh & Scheuren, 1978), and those that rely purely on the information contained in the survey data file. Although provisions for performing descriptive analyses on weighted data are available in standard statistical packages, these packages contain no formal procedures for performing the reweighting to deal with nonresponse. This presents no great problem in the case of weighting based on information in the survey file, because the programming to perform the reweighting is quite simple. An example of a program for reweighting in the SAS language is shown in Appendix B.

Among the single-valued explicit imputation methods, there are three alternative categories:

a. synthetic estimates, such as regression function values;

b. "hot deck" estimates, which assign a response taken from some other case on the file; and

c. distributional estimates, which assign a response randomly from an appropriately selected distribution.

Of these, the hot deck methods have received most attention recently. Synthetic estimates are available, however, in the BMDP system (Dixon & Brown, 1979), while the other two, "hot deck" and distributional estimates, have not been disseminated in common statistical packages. The procedure described in detail in this guidebook, PROC IMPUTE, is a distributional estimation method, embedded in the SAS package (Helwig & Council, 1979) for easy access.

Multi-valued explicit imputation, proposed by Rubin (1978), consists of imputing values several times, using different models of nonresponse and different random numbers, to create several copies of the file of data. Variance in the results of analyses among these files then provides an estimate of "error due to imputation." It has not been widely used because of its unpleasant requirement that all users of imputed data repeat all their analyses several times. This method may yet be proven to be necessary, however.

To decide among these methods, and to decide how best to plan ahead for imputation, survey designers and data analysts must consider the six questions stated above. We discuss each in turn.

(1) What resources are available for performing the imputation?

Imputation of missing data according to statistical models may require a complex computer program or a simple one, depending on the method used; unless a packaged procedure is available, writing programs for implementing the complex methods will require both a substantial programming effort and a clear understanding of the types of bias that imputation procedures can introduce.

There are three major statistical packages for handling survey data: BMDP, SPSS (Nie et al., 1975), and SAS. Numerous other packages are available at particular computer centers, and analysts should be familiar with provisions, if any, for imputing missing data at the computer centers

they commonly use. In BMDP, there is a program, BMDPAM, that is very easy to use and has five alternative methods: setting values to the mean, plus four regression estimates; using one variable, using two variables, using all available variables achieving statistical significance, and using all available variables. In SPSS, there is little that can currently be done with missing data. The regression and factor analysis routines in SPSS do, however, provide superficial methods for dealing with missing data in calculating residuals and factor scores. In SAS, a procedure, PROC IMPUTE, has been developed by SAGE under contract to NCES, that is very easy to use and at present has two alternative methods, regression subsetting and simple regression.

The cost of running either the BMDPAM or the SAS PROC IMPUTE program on a data file is on the order of magnitude of performing regression analyses on the file. The typical cost of runs on 20 variable files with 1000 cases on the NIH Computer Center IBM 370-168 system has been on the order of $10. With this guidebook (or with the BMDP manual), a programmer with SAS (or BMDP) should be able to set up a run within an hour.

(2) How big is the data file?

A survey data file has two dimensions of size: the number of variables and the number of cases. Each has substantial effects on the cost of imputation of missing data as well as on most other analyses. The number of cases affects the computer time required, and the number of variables affects both the time and storage required. Because imputation by PROC IMPUTE requires three passes through the file, compared to two passes for many other methods, it may be less attractive in its present form* for very large files (e.g., over 50,000 cases). Costs increase linearly with number of cases. For any method of imputation that makes use of relations among variables, the costs increase more than quadratically with the number of variables, however. If the file to be analyzed contains more

*All but the final pass through the data are for the purpose of parameter estimation, however, and could be run on a sample from very large files.

than 80 variables or so, it is advisable to impute variables in blocks of 50 to 80 each to limit costs. PROC IMPUTE can be called repeatedly on a file, with new variable lists, with no difficulty, so the only problem is to select blocks of variables appropriately. The recommended approach is to include variables that are highly related to each other in the same block. These relations can be determined either logically or on the basis of correlations. As a practical example, if imputation is performed on a file that is the merger of several years' surveys, then all years' values for any particular variable should be included in the same block because they will be highly related. To capture relations between blocks, variable-lists for successive blocks after the first should include key variables from earlier blocks for use in regression estimates.

From a theoretical perspective, it is also important to limit the number of variables in each block to a small fraction of the number of cases on the file (or to be more precise, the number of cases with data) to provide for stable estimation of parameters used in the imputation. The number of parameters estimated for use in imputation increases with the number of variables in each block. The number of parameters to be estimated can also be controlled by varying the coarseness or fineness of the imputation. PROC IMPUTE uses information about the size of the file obtained in the first pass through the data in order to determine the appropriate number of parameters--the fineness of the imputation--to estimate in the second pass through the data.

(3) What is the purpose for imputing missing data?

Imputation should be considered as but a step in a general plan for making use of survey data. It follows editing of the data, which should remove clearly spurious values from the file, so that they are not perpetuated by imputation and later analytical procedures. The selection of alternative imputation procedures depends on the uses to which the data are to be put. Several alternative purposes for imputation are shown in Table 1.

TABLE 1

PURPOSES FOR IMPUTATION

(USES OF DATA FILES)

1. TO ESTIMATE POPULATION TOTALS.

   --IMPUTATION IS FAIRLY EASY.

2. TO ESTIMATE RELATIONS AMONG MEASURES.

   --IMPUTATION MUST BE SOPHISTICATED.

3. TO TEST A COMPLEX SET OF HYPOTHESES.

   --IMPUTATION MUST BE SOPHISTICATED.

4. TO PRODUCE A "PUBLIC USE" FILE.

   --IMPUTATION MUST BE SOPHISTICATED;
   PARTICULAR UNITS SHOULD NOT BE IDENTIFIED.

5. TO MEASURE PARTICULAR UNITS (E.G., FOR AUDITS).

   --IMPUTATION IS NOT APPROPRIATE, UNLESS
   HIGHLY ACCURATE.

First, if the purpose is merely to estimate population means or totals, various methods work nearly equally well. Cases may be reweighted within strata, a simple "hot deck" procedure (within strata) can be used, or linear regression estimates will suffice. Linear regression estimates are available in the BMDP package as well as in PROC IMPUTE. To the extent that the distributions of respondents and nonrespondents overlap, these methods will produce accurate estimates (subject to assumptions described in answer to question #6). In fact, for this purpose, it is not even necessary to impute actual values; direct "macro-imputation" of totals based on summaries of relations between the presence of a variable with the values of other variables will suffice. (The term macro-imputation is used to refer to methods that can be implemented using only file summary data without requiring any additional examination of individual records on the file.)

To estimate relations among variables or to test complex hypotheses, the second and third purposes in Table 1, a more sophisticated method of imputation is necessary. This is the most common use of survey data in report generation. Relations may be presented as correlation coefficients, as graphs relating measures, as bivariate frequency tables, or as tables of means in different strata. The testing of complex hypotheses may go further to examine the factor structure of a set of measures or to compare mean differences to error estimates. In all these cases, imputation must not unduly distort the distributions of variables. Preservation of the multivariate distribution of variables is a problem not considered by most statisticians who are studying missing data imputation; it is, however, a primary goal of the development of PROC IMPUTE.

In particular, variances and covariances, as well as means, must be accurately reproduced in order to provide an analyzable file. Assignment of mean values, or even linear regression estimates, substantially reduces the variances of imputed variables; this problem is overcome, however, by procedures that assign values from distributions, such as "hot deck" procedures, and procedures that assign values randomly as distributed estimates, such as PROC IMPUTE. To preserve correlations among variables,

it is important to avoid imputing variables independently from each other. This is accomplished automatically by case reweighting methods and "hot deck" procedures that replace whole cases. Methods that impute variables one-by-one must use imputed values for predictor variables in imputing other variables in order to preserve correlations. Although it might appear that using imputed values to impute other values only builds error on error, the contrary is true when the purpose is to reproduce the multivariate structure of a data file rather than to make the best guess for each individual case.

PROC IMPUTE, unlike BMDPAM, assigns values as distributed random variables and uses imputed values in imputing other variables, and as a result it generally reproduces variances and correlations more accurately, although it reproduces individual values less accurately.*

If the purpose of imputation is to produce a "public use" file, the most sophisticated methods should be used. Because the analyses performed on a public use file cannot be predicted, tests of the validity of imputation (e.g., based on telephone follow-ups) are important to ensure that results of future analyses do not reflect imputation. Moreover, the method used should allow for easy estimation of the errors introduced when imputed values are included in subsequent analyses. Rubin (1978) has recommended producing replicate-files with different imputations so that users can perform replications of analyses to estimate the effects of variation in imputation. As he pointed out, imputed values will differ from recorded values both due to random error and due to errors in the assumptions underlying the model. By including explicit random error distributions in its calculations, PROC IMPUTE allows direct estimation of the random error component. This is described in Section II of the guidebook.

---

*This tradeoff appears to be unavoidable. The "SIMPLE" option in PROC IMPUTE allows it to mimic the performance of BMDPAM in reproducing individual values rather than variances and correlations.

Because a trade-off exists between reproducing individual values and reproducing distributions, one must frequently be sacrificed if the other is to be optimized.* Uses that require accuracy of individual values are those in which some future action is anticipated with respect to particular cases, such as stratified sampling for a future survey.

Because of the impossibility of complete elimination of error in individual cases, we recommend that imputed values not be used for purposes involving identification of individual cases. Imputation can then focus on reproducing distributions.

(4) What structures exist in the recorded values? Most surveys have internal logical structures, or redundancies, such as blanks for male, female, and total counts of staff. Imputation can, but should not be, undertaken blindly without cognizance of these structures. Whenever possible, constrained missing values should be filled in as a part of editing prior to imputation, to simplify the imputation task. For example, if male and female counts are present but the total is missing, the best method of filling in the total is obvious, but it will be different from the best method for use when all three counts are missing. The best method in the latter case might involve first estimating the total, then the components.

Analysts should, when possible, construct derived variables that indicate characteristics of the cases better than the basic survey response variables, such as teacher/pupil ratios for schools. Adding these variables to the file will increase the accuracy of imputation as well as of other analyses. On the other hand, to avoid bias, it is important not to impute values of variables ultimately to be used in analysis as nonlinear functions of other variables. For example, if one

---

*Imputing the appropriate modal value for all missing cases is optimal for the purpose of individual matching, but this will bias nearly all analyses.

imputes counts of teachers by multiplying imputed teacher/pupil ratios by counts of students, the resulting distribution of counts of teachers will be biased. Derived variables should be used as linear predictors in order not to introduce bias.

Imputation will obviously be more accurate when closer relations exist among variables present and those missing. Therefore, (1) if imputation must be done in blocks of variables, highly correlated variables should be included in the same block; and (2) if a high proportion of nonresponse to a particular item is expected for certain survey strata, then another item or items highly correlated with the target item but more likely to produce responses should be included in the survey instrument. An example of inclusion of a highly correlated simple variable would be a request for grades served by a school in addition to a grade-by-grade breakdown of enrollment.

(5) What is the pattern of missing data?

Six common patterns of missing data are shown in Figure 1. The recommendations for imputation vary between them. If data are missing randomly from the file, then imputation is only for convenience. Statistical computations based on the incomplete data file will, by definition, produce the same results that would have occurred had data not been missing, although the effective sample sizes are smaller. This situation is so rare that it need not be considered: respondents do differ from nonrespondents. For random variation, PROC IMPUTE is to be preferred over, for example, filling in mean values, because it reproduces distributions. For the case of attrition, when only a small amount of information (such as stratification-variable values) is known about nonrespondents, weighting is at least as good as other imputation methods, especially if the data file is already weighted, because no new complexities are introduced into the analyses. This situation is typical of one-time household surveys, where only the location of nonrespondents is known. When some variable is missing for all cases or is present for so few cases that stable parameters of its distribution cannot be obtained, then no

FIGURE 1

PATTERNS OF MISSING DATA

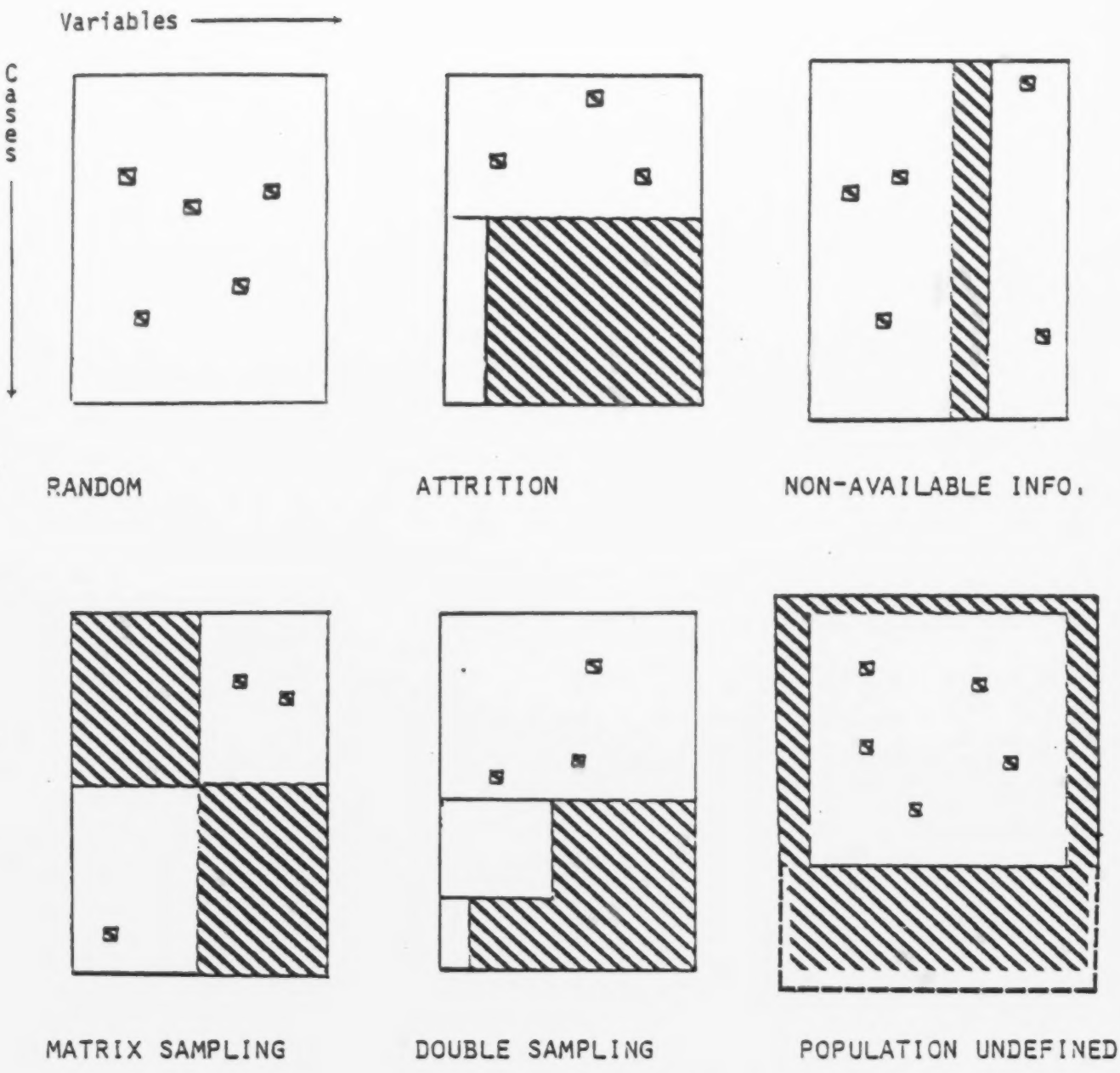


FIGURE 1. PATTERNS OF MISSING DATA. NOTE THAT DATA FILES ARE REPRESENTED AS RECTANGULAR MATRICES, WITH VARIABLES CONSIDERED AS COLUMNS AND CASES CONSIDERED AS ROWS. MISSING DATA ARE REPRESENTED BY DIAGONALLY FILLED AREAS.

imputation method is appropriate and the variable must be dropped from further analysis. This may occur, for example, when an intermediate aggregation agent, such as a State Education Agency, decides that no information on a particular measure should be reported for any school in the state.

The most common situation is one in which different blocks of variables are missing for cases of different "types." The types may be determined by the survey designer, for example, by following up nonrespondents using a shortened form of the survey instrument (e.g., a telephone follow-up of a mailed survey). They may also be determined by the respondents--respondents with particular characteristics may tend not to respond to certain items. This is the situation for which BMDPAM and PROC IMPUTE are most clearly useful. Weighting is an inefficient form of imputation in this situation because separate weights must be obtained for each variable.*

One other important pattern of missing data is unknown undercoverage of the universe. This will occur when the survey involves defining the universe as a combination of lists from numerous sources. One cannot always be sure that a sufficient set of sources has been checked to identify all members of a universe. If no information is known about nonrespondents, including their very existence, then no imputation method based on the survey data file alone is meaningful. An external source of data, known to represent the entire population, can be used, however, to impute missing values. This is commonly done by reweighting survey respondents so that their distributions on key variables match the distributions obtained from external sources (e.g., Oh and Scheuren, 1978).

*Cox and Folsom (1979) have proposed a method of variable by variable imputation that is mathematically equivalent to reweighting, but this method does not preserve relations among imputed variables.

(6) What assumptions are acceptable for the imputation? Every imputation method is based on a model of nonrespondents, a set of assumptions about what their responses would have been. Statistical analyses are only meaningful in terms of these models, so the model must be made explicit for any successful imputation procedure. All the models underlying methods that do not rely on external data are of the form: nonrespondents and respondents are alike, once particular differences are accounted for. The various methods differ in what types of differences they take into account, as shown in Table 2.

The assumption that relations among variables are constant is basic to nearly every imputation method. This is made explicit in regression-type methods, such as used by BMDPAM and PROC IMPUTE, but it is also present in all stratification weighting schemes and in hot-deck procedures that assign by strata or according to a nonrandom ordering of the file. The validity of the assumption of constant relations cannot be directly tested in practice, because data are not available on nonrespondents. An approximation can be obtained, however, by comparing relations across strata of respondents that differ in ways similar to respondent-nonrespondent differences.

A logical basis exists for the assumption that relations are constant even though respondents and nonrespondents may be quite different in level and variability of characteristics, and evidence exists to support this assumption. Further exploration of the assumption and the conditions under which it is satisfied are needed, however. The logic behind the assumption is that an observed relation is an observed invariance. Two variables cannot be highly correlated unless there is a combination of these variables that is nearly constant across the range of observations (e.g., counts of teachers and pupils are highly correlated because schools hold teacher/pupil ratios relatively invariant). For measures of level or variability, no similar invariance exists (other than finding that variability is near zero). Empirical results based on Project TALENT's special follow-up of 10,000 nonrespondents to its survey of 29-year-olds eleven years after high school graduation also support the assumption.

TABLE 2

MODELS (ASSUMPTIONS)

"RESPONDENTS AND NONRESPONDENTS ARE ALIKE, ONCE YOU TAKE INTO ACCOUNT...."

1. NO DIFFERENCES, ($Y_{RESP} = Y_{NONRESP}$)
   --IMPUTATION IS ONLY FOR CONVENIENCE.

2. $Y = f(X)$ IS INDEPENDENT OF RESPONSE/NONRESPONSE.
   A. Y IS A POINT VALUE OR A DISTRIBUTION.
   B. f IS AN EXPLICIT FUNCTION OR A SEARCH PROCEDURE.
   C. X IS ONE, TWO, A FEW, OR MANY DIMENSIONAL.
   D. X IS THE SAME FOR ALL Y'S OR DIFFERENT.

3. $Y = f(X)$ DEPENDS ON WHETHER THE CASE IS A RESPONDENT OR NONRESPONDENT (FOR VARIABLE Y).
   --IMPUTATION IS NEARLY IMPOSSIBLE AT PRESENT.

4. THE DISTRIBUTION OF Y IS KNOWN EXTERNALLY.
   --IMPUTATION BY "RAKING," OR REWEIGHTING.

Y DENOTES THE DISTRIBUTION OF VALUES OF A TARGET VARIABLE TO BE IMPUTED, X DENOTES THE VECTOR OF OTHER VARIABLES THAT PROVIDE INFORMATION ABOUT Y, AND F IS THE FUNCTION RELATING X TO Y.

Whereas nonrespondents differed substantially from respondents on measures obtained in high school and on later survey items, they did not differ significantly with respect to important relations. For example, although respondents had higher academic aptitude scores than nonrespondents and although the distribution of occupations differed between respondents and (followed-up) nonrespondents, the difference in academic aptitude between respondents and nonrespondents was generally the same across occupations.

To summarize, on the basis of these six questions, survey designers and data analysts should follow the flow diagram shown in Figure 2 in planning for, executing, interpreting, and using the results of imputation of missing data. Imputation must be planned prior to data collection. The most important consideration is to take steps to minimize nonresponse. For example, the survey instrument should be carefully pretested and edited; a sufficient rationale should be developed to convince individuals to respond, including letters of support from authorities; and a human relationship between the respondent and the person responsible for data collection should be established. In addition to minimizing nonresponse and planning for follow-up of nonrespondents, survey designers should search for related data to assist imputation. For example, Census data can be used to characterize the types of children attending a school district that fails to respond to an item on a survey instrument.

The flow diagram for imputation after data collection has three main paths, and we are primarily concerned with the choice to use PROC IMPUTE, the most common case. A key step in this process is the examination of the results of PROC IMPUTE to determine whether the imputation was sufficiently likely to be accurate. There are basically three conditions in which imputation can be adequate, in terms of matching distributions. First, if only a small amount of data is missing for a variable, imputation is not likely to affect analyses involving that variable greatly. Even if a large amount of data is missing for a variable, the imputation can be considered adequate if there is a strong relationship between the variable and other measures on the file. As described in Section II, one report generated by PROC IMPUTE contains estimates of the strength of

Figure 2

Flow Diagram for Imputation

(Prior to data collection)

Include simple items correlated with items for which nonresponse is expected.

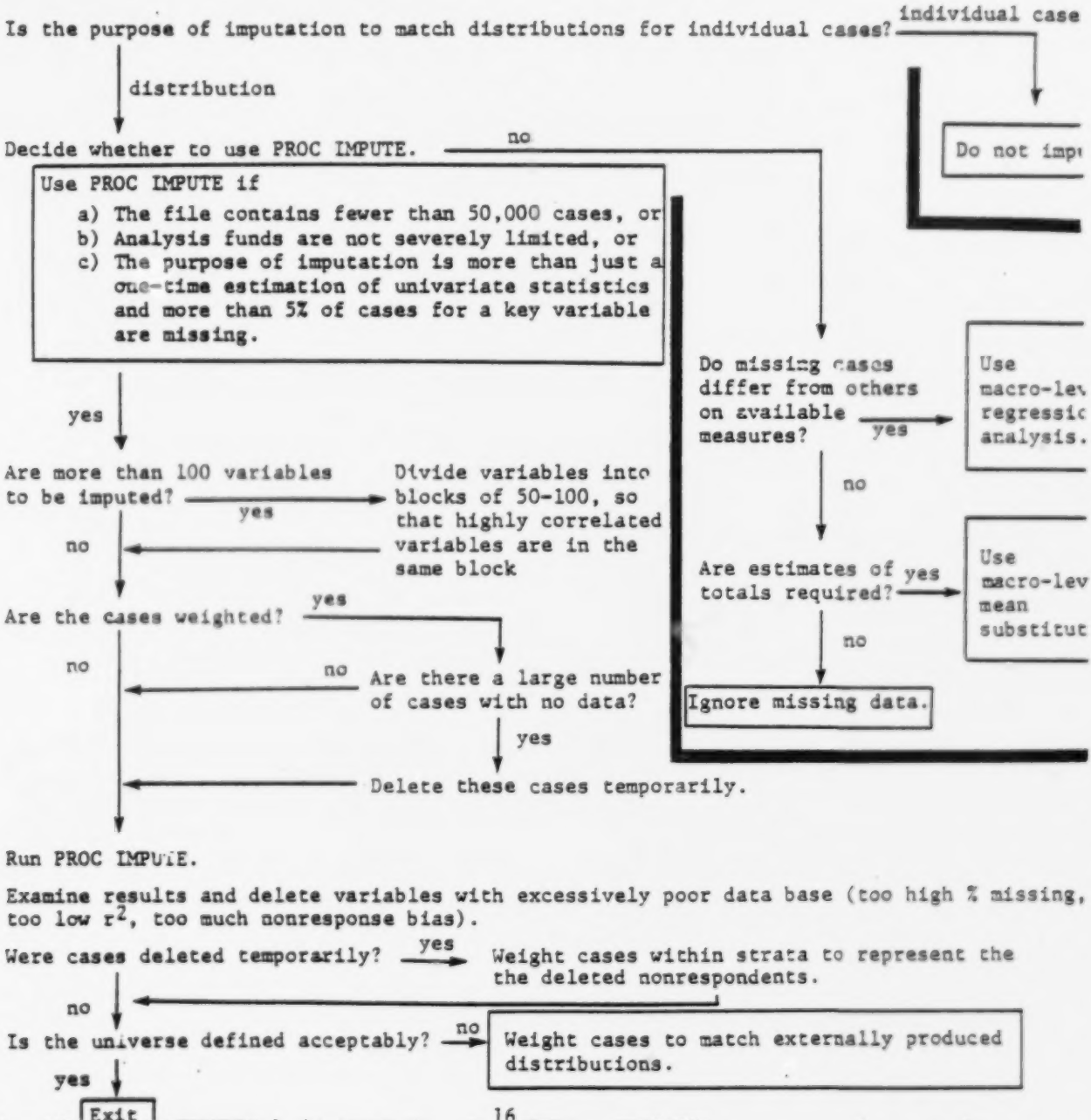
Take steps to minimize nonresponse.

Select a sample of cases for intensive follow-up, pending nonresponse.

Create a skeleton data file.

Merge external information onto the skeleton file.

(After data collection and editing)


Is the purpose of imputation to match distributions for individual cases?
individual case
Do not imp
distribution
Decide whether to use PROC IMPUTE.
no
Use PROC IMPUTE if
a) The file contains fewer than 50,000 cases, or
b) Analysis funds are not severely limited, or
c) The purpose of imputation is more than just a one-time estimation of univariate statistics and more than 5% of cases for a key variable are missing.
Do missing cases differ from others on available measures?
yes
Use macro-lev regressic analysis.
no
yes
Are more than 100 variables to be imputed?
yes
Divide variables into blocks of 50-100, so that highly correlated variables are in the same block
no
Are estimates of totals required?
yes
Use macro-lev mean substitut
no
yes
Are the cases weighted?
no
no
Are there a large number of cases with no data?
Ignore missing data.
yes
Delete these cases temporarily.
Run PROC IMPUTE.
Examine results and delete variables with excessively poor data base (too high % missing, too low r^2, too much nonresponse bias).
Were cases deleted temporarily?
yes
Weight cases within strata to represent the the deleted nonrespondents.
no
Is the universe defined acceptably?
no
Weight cases to match externally produced distributions.
yes
Exit

relations used for imputation. Finally, even if there is a reasonably large amount of missing data (e.g., 50%) and a fairly weak relationship (e.g., $r^2 = .25$), the imputation may adequately reproduce distributions if respondents do not differ from nonrespondents. One report generated by PROC IMPUTE displays the differences (on all other variables) between cases with a particular variable present or missing. Further work will be necessary to determine the appropriate combinations of these three conditions to use in making final decisions concerning acceptance or rejection of a particular variable's imputation.

II. THE NEW NCES ALGORITHM: PROC IMPUTE

After reviewing the options for imputation available in the common statistical packages, it was determined that something more was needed. The BMDPAM program in the BMDP series can satisfy a limited range of imputation needs, but the strong bias in variance and covariance estimates generated from the values imputed by BMDPAM left much to be desired. Other special purpose programs are not readily available and are inefficient to use because the user must devote considerable time to defining input and output formats and other parameters describing the data. Since this effort is already included in the use of the statistical packages for analyses, no extra effort is needed if an imputation procedure can be included within one of these packages.

It was decided to implement a new routine for missing data imputation in the Statistical Analysis System (SAS) because of the ease of implementing new routines in SAS, the great flexibility of this system for data manipulation, and the high level of use of this system. The use of SAS has increased dramatically over the past two years and now surpasses the use of SPSS or BMDP at most installations where it is available. (See recent NIH computer facility usage statistics for example).

The procedure implemented, PROC IMPUTE, is a distributional estimation procedure that is believed to be more general and to produce more accurate results than a standard "hot deck" procedure. Basically, this procedure considers each variable on the file in turn as a "target" variable whose missing values are to be filled in and uses information on other variables to minimize the error in imputing each target variable. For each "target" variable, regression analysis is used to find the best combination of predictors, and cases with the target variable present are divided into subsets based on values of the regression function. All cases in a given subset that are missing the target variable then have values assigned with random frequencies proportional to the distribution of reported values for that variable within the subset. The basic assumption of this algorithm is that <u>within these homogenous subsets</u>, the missing value cases will have

the same target value distribution as the cases with reported values on the target variable.

The following sections describe the PROC IMPUTE procedure more explicitly. The next section describes the algorithm in more detail. This is followed by sections that describe the steps necessary to run PROC IMPUTE and how to interpret the results of PROC IMPUTE. Time requirement estimates are given in Appendix A.

How PROC IMPUTE Works

PROC IMPUTE makes three passes through the input data file. The processing that occurs during and between each of these passes is described here in general terms to document the statistical algorithm. The specific input statements needed to run PROC IMPUTE and the output generated by PROC IMPUTE are described in later sections.

During the first pass through the data, basic univariate and bivariate statistics are computed. These include the mean, standard deviation, minimum, maximum, and number of missing values for each variable, the intercorrelations among the variables, and the number of cases missing one variable but not the other for each pair of variables (as well as pairwise means and standard deviations). Reports 1 through 3, described later, print out this basic information for the user.

Following the first pass through the data, stepwise regression analyses are performed "simultaneously" for each variable to be imputed. During these analyses, an ordered list of the imputation variables is constructed, and the regression analysis for each variable is limited to predictors that "precede" the target variable in the imputation list. The determination of the optimal ordering is a complex procedure based on relative amounts of missing data and the relative strengths of relations among variables. Initially no restrictions are imposed. Then, at each step, one predictor variable is added to one regression equation and additional restrictions are imposed by the fact that the new predictor is

forced to "precede" the target variable. The predictor-target pair selected at each step is that pair that will provide the greatest increment in the variance explained for all of the missing values (of all target variables). This process terminates when there are no more permissible predictors that provide a significant increase in the prediction of any of the target variables.

The predictions derived from these restricted regression equations may not be optimal. If variables X and Y are closely related, each should be used in the imputation of the other when possible. To allow for this, some variables must be imputed twice, considering the first imputation as a "ghost imputation" to be replaced later. Once the initial imputation list and the associated regression equations have been constructed, the imputation target variables are each reexamined (in their order in the imputation list). Additional regression equations are generated whenever the addition of "follower" variables would significantly improve the prediction.

Finally, for each regression equation, a number of subsets are defined in terms of regression function values. Within each subset, the distribution of target variable values can be expected to have a much smaller variance than overall, if the regression equation represents a strong relation. (The number of subsets is defined in terms of a trade-off between fine-grain-ness and stable parameter estimation. The number will vary with the expected number of cases with "complete data" for the regression equation variables.)

During the second pass through the data, regression function values are computed for each case and each equation where all the required variables are present, including the target variable. The complete bivariate frequency distributions of the regression function values and their associated target variables are estimated by counting the number of cases in each regression value subset at each level of the target variable. Following the second pass, each bivariate frequency distribution is converted to separate probability distributions for each regression subset. Figure 3 shows an illustration of these separate distributions.

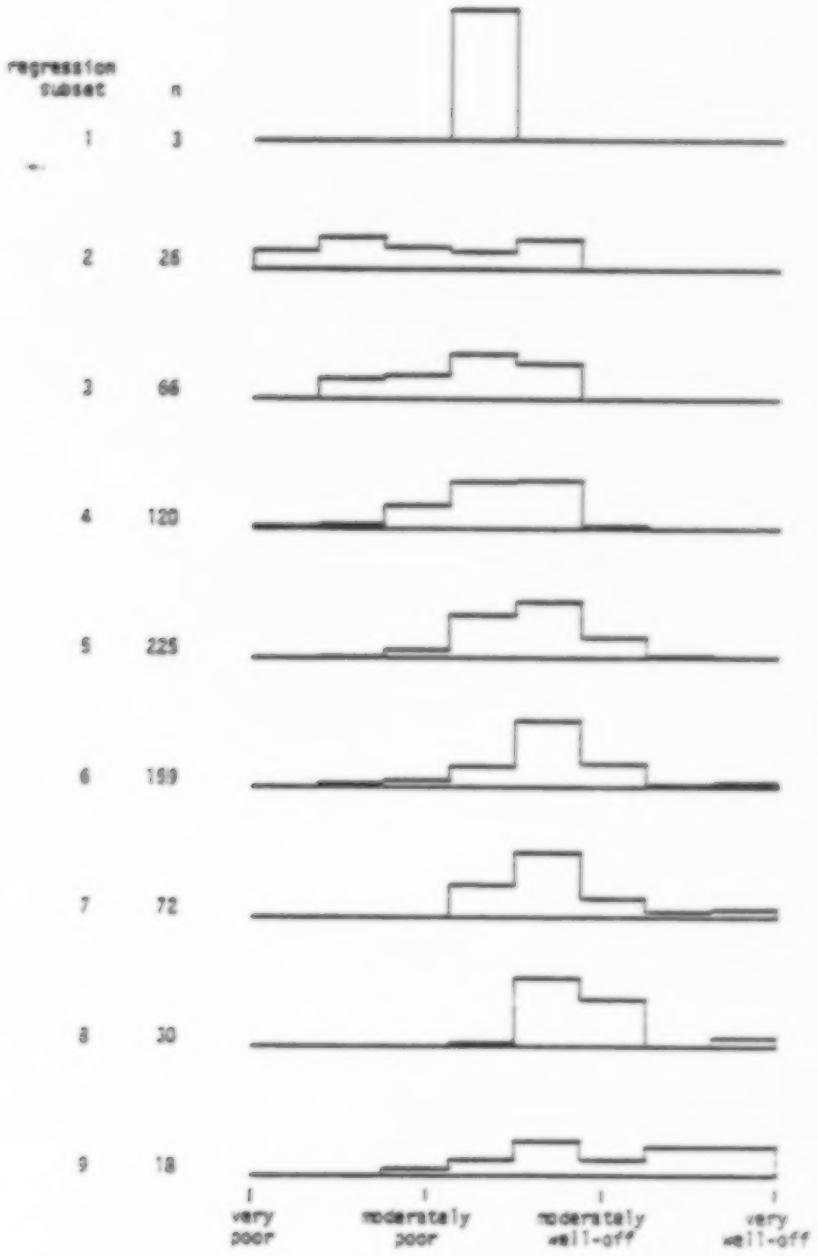


FIGURE 3. Distribution of target variable for each regression-function subset.

SCHOOL DISTRICT SES MEASURE
(School TV Utilization Study: NCES, 1979)

Note: The regression function was selected to account for maximum variance in the SES measure. Values were then partitioned into 9 discrete categories. The "n" refers to the number of cases in each regression-function category.

During the second pass through the data, the mean regression function value in each subset is also computed to provide information for interpolation between the distributions in adjacent regression subsets.

In the final pass through the data, missing values are imputed for each case. For each of the regression equations where a target value is missing, the regression function value is computed. The appropriate regression value subset and the adjacent subset are identified. A uniform pseudorandom variable between 0 and 1 is generated, and a value is computed for imputation of the target variable for each adjacent subset, based on the pseudorandom variable. The pseudorandom value is considered to be a probability, and the point on each cumulative distribution function (obtained in the second pass through the data) corresponding to that probability is identified (i.e., the inverse of the cumulative distribution function is applied to the random variable). If the "SIMPLE" option is specified, the pseudorandom variable is reset to .5 so that the median value for the subset is always selected. The imputed values obtained for the two adjacent subsets are then averaged according to the distance of the mean regression value in each subset from the regression value for the case being imputed. This average value is rounded to an integer if the integer flag is set for the target variable.

After all missing values have been imputed for a case, the case is written to the output file with all of the missing values filled in. Missing data flags are also created and set for each variable with a value of "I" corresponding to imputed values, a blank value for real values.

How to Use PROC IMPUTE

To use PROC IMPUTE, you must specify (1) the job control language (JCL) statements to execute SAS, to specify data sets, and to include the IMPUTE program in the standard SAS program library and (2) the SAS statements that call PROC IMPUTE. Figure 4 shows both kinds of statements for a sample run of PROC IMPUTE at the NIH Computer Facility. (PROC IMPUTE is currently being installed at the Data Management Center. The

FIGURE 4

SAMPLE JOB CONTROL CARDS

FOR RUNNING PROC IMPUTE ON A SAS

SYSTEM FILE (AT THE NIH COMPUTER CENTER)

```
//ACCTINIT JOB (ACCT, CLASS, TIME, LINES), USERNAME

// EXEC RUNSAS, REGION=400K

//LIBRARY DD DSN=WPG4T00.SAGELIB, UNIT=FILE,VOL=SER=FILE26,
//         DISP=SHR

//FT06F001 DD DUMMY

//OLDFILE  DD   DSN=YOUR OLD FILE, VOL=SER=YOUR VOLUME NUMBER,

//NEWFILE  DD   DSN=YOUR OLD FILE, VOL=SER=YOUR VOLUME NUMBER,

//SYSIN DD *

TITLE  YOUR RUN OF PROC IMPUTE:  (OPTIONAL)

PROC IMPUTE DATA = OLDFILE.SASNAME
                   OUT=NEWFILE.SASNAME;

     VAR (LIST OF VARIABLE TO BE PROCESSED.  IF OMITTED,
        ALL NUMERIC VARIABLES WILL BE PROCESSED);
```

library name and SAS procedure name will vary slightly, but otherwise the same statements will be required.) The remainder of this section describes each of the required statements more fully.

Job Control Statements

The JOB statement is the same as for any other run. See Appendix A for sample time estimates.

The EXEC statement uses the normal cataloged procedure for SAS.

The LIBRARY statement points to the SAGE library containing the program for PROC IMPUTE. The cataloged SAS procedure concatenates (adds) this library to the standard SAS library. In addition to the DSN (data set name), the UNIT (device type), VOLume (specific disc pack), and DISPosition (SHR for share) must be specified.

The FT06F001 DD (data definition) statement is required by some of the IMSL (International Mathematics and Statistics Library) subroutines that print warning messages. Since PROC IMPUTE reacts to these warnings itself, they need not be printed. The example shows how to specify a "dummy" output file for these warning messages.

The file data definition statements tell PROC IMPUTE the name and location of the input and output data files. If only PROC impute is run, these will be SAS system files. It is possible, however, to include other SAS statements to read and/or write raw data files and perform other analyses in the same run. If an output file is not specified, the imputed values will only be retained on a temporary file for use in the same run. Section 8 of the SAS Manual and the section on DD statements in the IBM JCL manual give complete information on the optional and required parameters associated with the data definition statements.

Finally, the SYSIN statement signals the beginning of the SAS statements.

SAS Statements

The PROC IMPUTE statement invokes the imputation procedure and provides key information for the imputation. SAS parses this statement using a "free field" format so that column positions do not matter. After the words "PROC IMPUTE," the parameters may be in any order. The following clauses may be included:

1. DATA=ddname.SASname points to the input data file. The "ddname" refers to the label used in the JCL. If omitted, a temporary SAS file is assumed. The SAS name is the internal data set name used by SAS. If no input data set is specified, the last data set created by SAS in this run is processed.

2. OUT=ddname.SASname points to the output data file. As before, ddname refers to a particular DD statement in the JCL, and SASname is the internal file name. If no ddname is specified, a temporary SAS file is created. If no output file data set is specified, a temporary data set is created using the standard SAS file default names.

3. SIMPLE is an optional keyword. If included, the SIMPLE option is invoked and the imputed values are all set to the median value of the target variable in the appropriate regression value subset. The use of this option is not recommended if there is any chance that variances and covariances will be analyzed. If this keyword is omitted, the default option is used and values are imputed randomly according to the target value distribution for the appropriate regression value subset. As with all SAS statements, the IMPUTE statement ends with a semicolon.

The VAR statement follows the PROC IMPUTE statement (possibly on the same line) and specifies the variables to be processed by PROC IMPUTE. If this statement is omitted, all numeric variables in the input data set will be processed. After the keyword VAR, the names of the variables to be processed are listed, separated by spaces. Only numeric variables may be included. The order of the variables in the VAR statement determines their order in the first three reports and also corresponds to the numbering of the missing data flag variables (MFLAGn) generated by PROC IMPUTE. The processing time and storage requirements depend primarily on

the number of variables included in this statement. (See Appendix A for examples.) The VAR statement ends with a semicolon.

The Output Data Set

In addition to the missing data reports described in the next section, PROC IMPUTE generates an output data set. The output data set is a standard SAS systems file and includes each of the variables specified in the VAR list plus a missing data flag variable for each of the variables in the VAR list. The missing data flag variables have names MFLAG1 to MFLAGn where n is the number of variables processed. For example, the statements "PROC IMPUTE; VAR X Y Z;" would produce an output file containing six variables: X, Y, Z, MFLAG1, MFLAG2, and MFLAG3. MFLAG1 would be set to the value "I" for all records in which X was imputed, and to the value " " (blank) for all records in which X was already on the file. Similarly, MFLAG2 and MFLAG3 would indicate whether Y and Z were imputed or actual values. The flags are character variables of length 1. These flag variables may be given new names by attaching a RENAME statement to the output data set specification in the impute statement. For example, "PROC IMPUTE OUT=DSKOUT.MYFILE (RENAME=(MFLAG1=MVAR1 MFLAG2=MVAR2 MFLAG3=MVAR3);" would assign the names MVAR1, MVAR2, and MVAR3 to three missing data flags. (See the SAS manual for further information on renaming variables.)

Not all variables on the input data set need be included in the VAR list for PROC IMPUTE; any variables not in the VAR list will not be on the output data set of PROC INPUTE. To combine the imputed values with the other variables not included in the VAR list, it is sufficient to execute the following SAS MERGE.

```
DATA MERGEDOUT;
     MERGE OLDFILE NEWFILE;
```

No "BY" statement is necessary because the file containing imputed values, NEWFILE, is a record-by-record transformation of the original data set, OLDFILE. If variables are imputed in blocks (e.g., 200 variables imputed

in four blocks of 50), a MERGE must be inserted after each call to PROC IMPUTE if some variables imputed in each block are used in imputing variables in other blocks.

Limitations of PROC IMPUTE

The following limitations apply to Version 1 of PROC IMPUTE. Some of these limitations will be removed in subsequent versions.

1. Only numeric variables can be processed. Character variables must be recoded prior to PROC IMPUTE if they are to be imputed.

2. Categorical variables are treated as if they were ordered, in the derivation of regression equations and subsets. This may not lead to an optimal set of predictors for these variables or to their optimal use in predicting other variables. It may be desirable to recode categorical variables into a series of dichotomous indicators prior to using PROC IMPUTE.

   For example, a school might be either "for girls only," "for boys only," or "for both boys and girls," coded "1," "2," "3." In this case, two dichotomies that might be useful in prediction would be (1) to combine "for girls only" with "for boys only," as opposed to coeducational and (2) to combine "for boys only" with "for both boys and girls," as opposed to schools not for boys. In this case, the original three-valued variable could easily be reconstructed from imputed values on the two dichotomies. Note that although it is theoretically possible for the imputation to produce conflicting values for the dichotomies, these cases should be very rare because no conflicts exist in the observed data and because one of the two dichotomies will almost certainly play a strong role in the imputation of the other. Nevertheless, the coding to reconstruct a categorical variable from dichotomies must handle possible conflicts.

3. Many survey instruments are designed so that certain items determine whether other items are to be skipped or not (e.g., respondents who did not attend college are not asked to indicate a college major). This version of PROC IMPUTE does not include a provision for indicating that certain values are to remain missing after imputation. There are basically two methods for handling "skip patterns" with the current version of PROC IMPUTE. (1) The file containing imputed values can be re-edited to set appropriately skipped items back to "missing." Alternatively, (2) variables conditional on a particular item can be imputed in a separate block, after the conditioning item has been imputed, and only for the subfile of cases for which the variables should be imputed. Because it is less expensive to make a series of calls to PROC IMPUTE on small blocks of variables than a single call on a large number of variables, it is advisable to handle a complex skip pattern through a series of calls to PROC IMPUTE on appropriate subfiles. The SAS system greatly facilitates the file manipulation (extraction of cases and later merging) needed for this.

4. Case weights are not used in the estimation of the imputation parameters. By including the weight variable in the variable list, however, it is possible to eliminate any first-order (but not interaction) effects associated with differential case weights.

5. While the number of variables processed by PROC IMPUTE is theoretically unlimited, the storage and processing time requirements (i.e., costs) increase dramatically for larger numbers of variables (over 50 or so).

Reports Generated by PROC IMPUTE

Missing Data Report #1: Missing Data Frequencies and Univariate Frequencies

Figure 5 shows an example of the first report generated by PROC IMPUTE. This report provides information on the amount of missing data and on the basic univariate characteristics of each variable. Specifically, the following information is provided:

Column	Description
1	Variable name. The variables are processed in the order specified by the VAR list. The order of the variables is important because the missing data flags are numbered in this order. If no VAR list is specified, all numeric variables are selected according to their position in the file.
2	The number of cases with missing values for this variable. Note that the specification of missing values is part of the work inherent in the creation of a SAS systems file. See Chapter 6 of the SAS Manual.
3	The percent of cases with missing values for this variable.
4	The number of cases with valid values for this variable.
5,6	The minimum and maximum reported values. Imputed values will always lie within the range of the reported values. When continuous or many-valued discrete variables are sliced into a smaller number of distinct levels, the minimum and maximum values are used as endpoints of the lowest and highest levels respectively.*
7	The integer/decimal flag. During input, the reported values are checked to see whether any noninteger values are present. If all reported values are integers, the variable is flagged an "integer" and all imputed values will be integers. If any of the reported values are nonintegers, the variable is flagged as "decimal" and noninteger values will be imputed.

* In Version 1A of PROC IMPUTE, available in August 1980, the minimum for many-valued discrete variables may print out as a very small positive number instead of zero. This is of no real consequence, but will be corrected in Version 2.

STATISTICAL ANALYSIS SYSTEM 22:49 TUESDAY, JULY 22, 1988

MISSING DATA REPORT #1: MISSING DATA FREQUENCIES AND UNIVARIATE STATISTICS

① VARIABLE	② N MISS	③ % MISS	④ N PRES	⑤ MIN	⑥ MAX	⑦ I/D	⑧ NV	⑨ MEAN	⑩ STD	⑪ LABEL
SD75IAFF	1167	58.53	827	1.3E-85	65	I	6	3.07739	6.43233	
SD11PRAC	312	15.65	1682	1	5	I	5	4.00416	0.625922	
SFACT1	795	39.87	1199	-3.4624	2.46973	D	7	.0579118	0.976573	
SFACT2	795	39.87	1199	-2.81099	3.55029	D	7	-.018764	0.988237	
SFACT3	795	39.87	1199	-5.49487	2.3686	D	7	.0269737	1.00111	
SFACT4	795	39.87	1199	-2.28418	4.29736	D	7	-.051538	0.982089	
SH1ENROL	191	9.58	1803	172	596653	I	8	17851.9	42738.7	
SH4SES	440	22.07	1554	1.24	3.99	D	8	2.64882	0.397517	
SH5ENGZL	248	12.44	1746	1.3E-85	100	I	8	6.8173	17.7858	
TMGT	0	0.0	1994	0.661	279.674	D	9	9.27072	11.6486	
TA3AHOST	66	3.31	1928	4	240	I	9	73.8511	51.6864	
TA3BHOCL	85	4.26	1909	1	14	I	14	3.20168	2.07978	
TB4AVHDD	641	32.15	1353	1	4	I	4	1.76349	0.837197	SET NOT AVAILABLE
TB6RECPT	660	33.10	1334	1	3	I	3	1.50375	0.632385	POOR RECEPTION
TC27677	12	0.60	1982	1	5	I	5	3.60494	1.54575	
TC3PGSER	672	33.70	1322	1.3E-85	5	I	6	1.40545	1.46999	
TC4AVEHK	26	1.30	1968	1	13	I	13	2.11026	1.81332	
TE111VCO	544	27.28	1450	1.3E-85	1	I	2	0.684138	0.464858	
TE8ADM	561	28.13	1433	2	5	I	4	3.57711	0.648421	
TF1TRAIN	49	2.46	1945	1.3E-85	1	I	2	0.188689	0.391261	
TC1V1	655	32.85	1339	1	3	I	3	2.76326	0.478133	
TC1V2	655	32.85	1339	1	3	I	3	1.98282	0.344166	
TFACT1	829	41.57	1165	-2.54614	3.9563	D	7	-.047753	0.968793	
TFACT2	829	41.57	1165	-2.86958	2.72339	D	7	-.001284	0.995243	
TFACT3	829	41.57	1165	-6.81148	1.72437	D	7	-.008924	1.00298	
TFACT4	829	41.57	1165	-4.14185	2.68449	D	7	.0025682	0.97626	
TI1ENROL	214	10.73	1780	0	995	I	8	39.7416	96.8167	

FIGURE 5. PROC IMPUTE Report #1, Example

8 The number of distinct levels to be used in approximating distributions for this variable. The program selects an optimal number of levels based on the number of cases with reported values. A greater number of levels is selected when more cases are available to use in estimation. If the variable is flagged as integer, however, and the actual range does not exceed twice the optimal number of values, then the number of integers in the range is used; otherwise, the optimal number of values is used.

9,10 The mean and standard deviation of the reported values are shown. The statistics are used later to generate raw regression coefficients. They are presented here to allow users to check the reasonableness of their input and as a reference for later information.

11 Variable labels, if present, are shown to aid in the identification of each variable.

Missing Data Report #2: Characteristics of Cases with Missing Values

Figure 6 shows an example of the second missing data report. This report summarizes the information that is available on cases with missing values. Each row of this report focuses on cases with missing values for a particular variable. Each column (except for diagonal cells) presents information on the missing value cases with respect to a particular other variable. For example, information in column 2 (PA5TSTAT) and row 1 (variable PA3TLADA) indicates how cases with and without PA3TLADA missing differ in terms of PA5TSTAT.

The first two entries in each cell give the mean and standard deviation of the column variable for cases with missing values on the row variable. For example, the mean value of PA5TSTAT for cases missing PA3TLADA is 34.3032, compared to an overall mean (shown in the diagonal cell) of 34.2886. In general, the column variable will be present for only a portion of the cases that are missing the row variable. The third entry gives the number of cases missing the row variable but not the column variable. For example, 432 cases were missing PA3TLADA but not PA5TSTAT. The fourth entry is the phi coefficient describing the correlation of the presence of data on the row variable with the presence of data on the column variable. (Note: Under Version 1A, the phi coefficient is incorrectly computed and should be ignored). The fifth entry in each cell gives a t-statistic measuring the extent of the column variable

MISSING DATA REPORT #2: CHARACTERISTICS OF CASES WITH MISSING VALUES (FOR EACH VARIABLE)

MISSING VARIABLE	SECOND VARIABLE PA3TLADA	PA3TSTAT	PA6SET	PB4CTLTV	PB14ATEA	PB14BITV	PC1USE	PC3COTR	PC9DIST	PFACT1	PFACT2
PA3TLADA	752.748	34.3832	1.82456	8.40858	48.3521	24.8184	4.86897	0.533333	3.75758	-.086234	0.215319
	556.847	39.0778	0.797426	10.935	62.9163	32.856	0.59293	0.498888	0.736634	0.987698	1.02144
	1383	432	114	342	514	494	348	60	429	93	93
	0.0	0.0	0.0	0.0	0.0	0.928	12.862	12.819	4.813	0.0	0.0
	0.0	0.818	-1.213	0.429	1.548	-7.518	4.237	-2.599	-1.234	-0.858	2.168
	1.0000	0.9923	0.2259	0.6681	0.1247	0.0000	0.0000	0.0095	0.2182	0.9593	0.8311
PA3TSTAT	985.479	34.2886	2.18127	7.16	59	22.6881	3.98182	0.723404	3.66667	0.187252	.0113647
	789.784	29.4854	0.858836	7.91672	112.735	26.462	0.632194	0.447315	0.620342	1.18154	1.10099
	48	1767	79	100	136	122	110	47	123	61	61
	0.0	0.0	0.0	0.0	0.0	12.603	0.0	0.0	0.0	3.991	3.991
	2.328	0.0	2.825	-1.316	2.438	-4.867	0.565	0.467	-2.355	0.726	0.119
	0.0202	1.0000	0.0434	0.1888	0.0155	0.0000	0.5721	0.6488	0.0191	0.4681	0.9054
PA6SET	985.667	35.0568	1.91232	8.26586	61.1262	21.0238	4.89274	1	3.7234	-.566087	0.956117
	222.66	41.8489	0.874108	11.6151	68.1131	31.1032	0.584983	0	0.779349	0.494032	1.12547
	3	352	1494	249	420	404	248	4	329	2	2
	0.0	0.0	0.0	0.0	0.776	1.178	0.0	23.545	0.0	0.0	0.0
	1.184	0.412	0.0	0.124	1.511	-9.593	4.179	19.379	-1.983	-1.563	1.206
	0.2368	0.6810	1.0000	0.9010	0.1320	0.0000	0.0000	0.0000	0.0580	0.1186	0.2282
PB4CTLTV	737.543	32.383	1.93038	8.1836	46.2103	7.43817	4.82273	0.833333	3.4876	-.082486	-.082034
	482.44	19.8325	0.908893	9.50463	89.4594	19.7292	0.583432	0.372678	0.824378	1.03116	1.09174
	140	282	158	1585	214	201	44	24	121	127	127
	0.0	0.0	0.0	0.0	10.998	0.0	0.0	0.0	0.0	8.174	8.174
	-0.386	-1.589	0.267	0.0	1.666	-18.581	0.849	1.847	-4.308	-0.036	0.032
	0.6995	0.1135	0.7892	1.0000	0.0969	0.0000	0.3965	0.0650	0.0000	0.9709	0.9748
PB14ATEA	811.547	40.1576	1.93182	7.78689	37.2048	N/A	3.94444	0.666667	3.88889	.0922983	.0214815
	549.999	31.5389	0.914473	9.79648	42.3275	N/A	0.650241	0.471405	0.657342	0.983372	1.04913
	159	165	176	61	1738	0	54	39	54	139	139
	0.0	0.0	0.776	10.998	0.0	0.0	8.484	0.0	0.0	1.387	1.387
	1.431	2.528	0.303	-8.323	0.0	0.0	-0.853	-0.368	1.078	1.159	0.315
	0.1538	0.0122	0.7628	0.7469	1.0000	1.0000	0.9581	0.7131	0.2821	0.2467	0.7527
PB14BITV	824.03	39.2837	1.99187	7.10714	53.4668	34.0602	4.87619	0.724638	3.81579	0.18162	-.014931
	564.572	29.3741	0.906283	7.73946	36.1326	32.7399	0.612493	0.446697	0.720207	0.953228	0.988091
	203	215	224	112	64	9674	185	69	114	185	185
	0.928	12.603	1.178	0.0	9.317	0.0	7.238	4.196	0.0	2.271	2.271
	1.953	2.661	1.428	-1.498	3.644	0.0	2.282	0.597	0.336	1.539	-0.148
	0.0513	0.0084	0.1564	0.1348	0.0003	1.0000	0.0283	0.5585	0.7373	0.1242	0.8822
PC1USE	708.22	31.3532	1.81343	4.85714	44.5435	1.36807	3.949	0.777778	3.33333	.0673684	.0440234
	473.587	18.625	0.890815	4.87252	95.9341	9.59433	0.689736	0.41574	0.871698	0.948255	1.12585
	123	269	134	21	184	171	1608	9	93	184	184
	12.862	0.0	0.0	0.0	8.484	7.238	0.0	6.795	14.363	0.0	0.0
	-1.073	-2.496	-1.349	-3.892	1.154	-32.757	0.0	0.689	-5.315	0.243	0.467
	0.2838	0.0134	0.1778	0.0021	0.2497	0.0	1.0000	0.5428	0.0000	0.4579	0.6498

ENTRIES ARE: MEAN/STD OF 2ND VAR FOR CASES MISSING 1ST VAR
NUMBER OF CASES MISSING 1ST VAR WITH VALUES FOR 2ND VAR
PHI CORRELATION OF MDFLAGS
T/SIG OF DIFF IN 2ND VAR BETWEEN CASES WITH & W/O 1ST VAR

Note. N/A indicated that the statistics could not be computed because no data fit the constraint (presence of column variable, but missing the row variable).

FIGURE 6. PRØC IMPUTE Report #2

mean difference between cases missing the row variable and the sample as a whole. The sixth and final entry in each cell gives the significance of this t-statistic. For example, the t-statistic comparing values of PA5TSTAT between cases with and without values on PA3TLADA is 0.010. This statistic is important for evaluation of the confidence that should be placed in imputations. Where substantial differences exist, the likelihood of deviation from the assumptions of the model are increased. Therefore, variables with substantial differences on key variables (in the estimation of the survey analyst) should be examined to evaluate (1) whether their missing data are so frequent and (2) whether their imputations are so poor, in terms of error variance, that the variables should be deleted from further analyses.

The first three entries in each cell can be compared with the corresponding entries in the column's diagonal cell. The diagonal cells give the means, standard deviations, and number of all reported values for the column variables. Comparison of the values in each cell with the values in the diagonal cell for that column indicates the extent to which the cases missing the row variable differ from the sample as a whole, at least insofar as that can be known given that the column variable may itself have missing values.

The information presented in this report is helpful in understanding the nature of the missing data in a particular survey. To the extent that these results indicate more frequent nonresponse for particular types of cases, it may be possible to modify future data collection procedures to decrease the nonresponse and omit rates for these cases. (For example, if the omit rate for some items were closely related to the respondent's reading ability, it might be possible to decrease this omit rate by simplifying the wording of these items.)

Missing Data Report #3: Correlations between Reported Values

Figure 7 shows an example of the third report generated by PROC IMPUTE. This report shows the correlations between each pair of variables based on all cases for which both variables are present. The number of

MISSING DATA REPORT #3: CORRELATIONS BETWEEN REPORTED VALUES

FIRST VARIABLE	SECOND VARIABLE PA3LADA	PA3STAT	PA6SET	PB4CTLTV	PB14AIEA	PB14BITV	PC1USE	PC3COIR	PC9DIST	PFACT1	PFACT2
PA3LADA	1.0000 1383 0.0	0.0984 1335 0.0	0.4895 1388 0.0000	0.2329 1243 0.0000	0.9392 1224 0.0	-0.2120 1188 0.0000	0.0546 1260 0.0525	0.2868 789 0.0000	0.0179 1256 0.5256	0.1185 1237 0.0001	-0.1024 1237 0.0003
PA3STAT	0.0984 1335 0.0	1.0000 1767 0.0	0.4934 1415 0.0000	0.3122 1485 0.0000	0.0878 1402 0.0	-0.1662 1552 0.0000	0.0388 1498 0.2335	0.2927 882 0.0000	0.0887 1562 0.7313	0.1873 1269 0.0001	-0.0798 1269 0.0044
PA6SET	0.4895 1388 0.0000	0.4934 1415 0.0000	1.0000 1494 0.0	0.8245 1336 0.3708	0.4766 1318 0.0000	-0.2681 1270 0.0000	0.0743 1568 0.0041	0.1845 845 0.0000	0.0526 1356 0.2389	0.0991 1328 0.0003	-0.0338 1328 0.2182
PB4CTLTV	0.2329 1243 0.0000	0.3122 1485 0.0000	0.8245 1336 0.3708	1.0000 1585 0.0	0.2258 1329 0.0000	0.2495 1473 0.0000	0.0924 1564 0.0003	0.0880 825 0.0184	0.1229 1564 0.0000	0.0289 1283 0.4688	-0.0613 1283 0.0336
PB14AIEA	0.9392 1224 0.0	0.0878 1402 0.0	0.4766 1318 0.0000	0.2258 1329 0.0000	1.0000 1738 0.0	-0.1617 1674 0.0000	0.0541 1554 0.0330	0.2987 810 0.0000	-0.0181 1631 0.4647	0.1181 1191 0.0001	-0.1232 1191 0.0000
PB14BITV	-0.2120 1188 0.0000	-0.1662 1552 0.0000	-0.2681 1270 0.0000	0.2495 1473 0.0000	-0.1617 1674 0.0000	1.0000 1674 0.0	0.3273 1503 0.0000	-0.2245 788 0.0000	0.3682 1571 0.0000	0.0384 1145 0.1960	0.0883 1145 0.0028
PC1USE	0.0546 1260 0.0525	0.0388 1498 0.2335	0.0743 1568 0.0041	0.0924 1564 0.0003	0.0541 1554 0.0330	0.3273 1503 0.0000	1.0000 1608 0.0	0.0941 840 0.0064	0.5846 1592 0.0	0.2991 1226 0.0000	0.0221 1226 0.4389
PC3COIR	0.2868 789 0.0000	0.2927 882 0.0000	0.1845 845 0.0000	0.0880 825 0.0184	0.2987 810 0.0000	-0.2245 788 0.0000	0.0941 840 0.0064	1.0000 849 0.0	-0.0437 846 0.2046	0.1554 776 0.0000	-0.0793 776 0.0273
PC9DIST	0.0179 1256 0.5256	0.0887 1562 0.7313	0.0526 1356 0.2389	0.1229 1564 0.0000	-0.0181 1631 0.4647	0.3682 1571 0.0000	0.5846 1592 0.0	-0.0437 846 0.2046	1.0000 1685 0.0	0.1542 1228 0.0000	0.0475 1228 0.0964
PFACT1	0.1185 1237 0.0001	0.1873 1269 0.0001	0.0991 1328 0.0003	0.0289 1283 0.4688	0.1181 1191 0.0001	0.0384 1145 0.1960	0.2991 1226 0.0000	0.1554 776 0.0000	0.1542 1228 0.0000	1.0000 1338 0.0	-0.0138 1338 0.6153
PFACT2	-0.1024 1237 0.0003	-0.0798 1269 0.0044	-0.0338 1328 0.2182	-0.0613 1283 0.0336	-0.1232 1191 0.0000	0.0883 1145 0.0028	0.0221 1226 0.4389	-0.0793 776 0.0273	0.0475 1228 0.0964	-0.0138 1338 0.6153	1.0000 1338 0.0
PFACT3	-0.0223 1237 0.4338	-0.0895 1269 0.7347	0.0250 1328 0.3813	0.0055 1283 0.8682	-0.0263 1191 0.3653	0.1076 1145 0.0001	0.2075 1226 0.0000	-0.0812 776 0.0727	0.1513 1228 0.0000	0.0184 1338 0.7047	0.0038 1338 0.8897
PFACT4	0.0877 1237 0.7875	-0.0198 1269 0.4984	0.0208 1328 0.4987	-0.0551 1283 0.0568	0.0098 1191 0.7369	-0.1673 1145 0.0000	-0.1969 1226 0.0000	-0.0641 776 0.0763	-0.0782 1228 0.0061	0.0251 1338 0.3687	0.0061 1338 0.8253

ENTRIES ARE: CORR/N/SIG

FIGURE 7. PRØC IMPUTE Report #3

cases with both variables and the significance level of the correlations are also printed. These correlations provide the basis for estimating prediction equations for imputation. The information presented in this report is virtually identical to the information printed by the SAS routine PROC CORR. It is presented here to eliminate the need for a separate run of PROC CORR. (Version 2 of PROC IMPUTE will include an option for omitting this report if PROC CORR has already been run.)

Missing Data Report #4: Regression Equations

Figure 8 shows an example of the fourth report generated by PROC IMPUTE. This report shows the regression equations used with each variable to be imputed (target variable). Regression functions are generated only for variables with some missing data. These variables are ordered so as to maximize the total variance accounted for in the prediction of all missing values when each variable is predicted only from preceding variables. This ordering is necessary to ensure that any missing values among the predictor values will have already been filled in before the variable is used as a predictor in a regression function. (Variables with no missing values are placed at the beginning of the list, thus they precede all of the variables to be imputed.)

After an equation has been generated for each variable to be imputed, each of the variables in the list is reexamined to see if its prediction could be significantly improved by including "follower" variables in the prediction equation. If so, a second equation is generated, and both equations will appear in Report #4. The variable will then be imputed a second time after an initial ("ghost") imputation has been performed for each of the missing values.

The leftmost columns show the target variable for each equation and an estimate of the squared multiple correlation, which is the proportion of variance of the target variable accounted for by the predictor variables. The actual variance accounted for may differ somewhat from the estimate shown here because:

MISSING DATA REPORT #4: REGRESSION EQUATIONS FOR EACH VARIABLE

DEPENDENT VARIABLE	MULT R2	PREDICTOR VARIABLES	STD COEF	RAW COEF	VARIABLE NAME	PART COV	VARIABLE NAME	PART COV	VARIABLE NAME	PART COV	VARIABLE NAME	PART COV
TA3ANOST	0.410	PA3TLADA	0.2834	.001483	TA3BNOCL	.420988	TA1G13	-.15871	PA6SET	.1541/4	TA4D7	-.11624
		TA1G912	0.2450	2.11549	IC6AVEMK	-.07429	IC3RGSER	-.06705	TA4D6	.065643	TA4D2	-.06522
		TA1G46	.062048		SA3ADA	-.06198	PB14ATEA	-.05993	SC2B78	-.05778		
		TA1G79	0.3861	3.34351	IC27677	.055984	SB5ATEA	-.05262	SC3IIVEX	.052289	SB3ILIV	-.05137
		CONST		1.1692	SC2C912	-.04985	SH1EMROL	-.04684	IB6RECPT	.040841	SH4SES	-.03677
					SD7STAFF	.033476	PB14BITV	-.03288	II4SES	.032199	TA4D4	-.03166
					SFACT2	.028868	SC43YRS	-.02732	IB6AVHDD	.026128	PFACT1	.023979
					THGI	-.02356	PB4CILTV	.023539	TA4D3	-.02334	PA5ISTAT	-.02205
					PG2DROPS	-.02143	IFACT2	-.02134	PC3C01R	.021192	PG4SES	.019355
					SD11PRAC	0.01893	IFACT4	-.01836	SD1RESP	-.01763	IE8ADM1	-.01552
					PG1ENROL	-.01482	PC9DIST	-.01363	IE11IVC0	-.01305	PG5ENG2L	-.01293
					PFACT4	.012617	SB5BITV	-.01227	SFACT4	-.01207	PFACT3	.011744
					IC1V2	.010892	IC1V1	-.01001	IFACT3	-0.0081	IF11BRAIN	.006543
					TA4D1	-.00637	PFACT2	.006203	II1ENROL	.005637	SFACT3	.005621
					IFACT1	.005215	SH5ENG2L	.004457	SC11LEXP	.003358	TI5ENG2L	.001979
					TA4D5	-.00185	PC1USE	-0.0012	SFACT1	-7E-04		
TA3BNOCL	0.726	TA3ANOST	0.8525	.038136	TA4D6	-.10506	PA6SET	.097338	TA1G79	.090269	TA1G912	.090159
		CONST		-.15254	PB14BITV	-.07866	TA1G13	-.06922	IC27677	.052671	SD7STAFF	-.04927
					TA4D5	.046854	SB5ATEA	-.04219	PFACT2	-.04185	IFACT2	-.04129
					TA4D2	.037431	SA3ADA	-.03986	IFACT4	.038669	IC3RGSER	-.03854
					IE11IVC0	.037155	SH1EMROL	-.03557	PG5ENG2L	-.03476	II1ENROL	-.03363
					II5ENG2L	-.03298	PB14ATEA	.032722	IC6AVEMK	-.03265	SC43YRS	.032329
					PFACT4	.032066	PA5ISTAT	.029599	TA4D1	-.02909	SFACT4	.028719
					SB3ILIV	-.02814	PA3TLADA	.027261	PC3C01R	.026982	PB4CILTV	-.02501
					SD1RESP	-.02491	IF11BRAIN	.024819	SC2C912	.024501	PG4SES	.022723
					TA4D3	-.01973	SFACT2	-.01768	SH5ENG2L	-.01736	SD11PRAC	.016577
					PFACT3	.016521	TA4D7	.016062	SB5BITV	.015614	IC1V1	.015586
					PG2DROPS	-0.013	SH4SES	.012371	IB6AVHDD	.011861	TA1G46	.011206
					PFACT1	0.01107	SFACT1	.010781	SC11LEXP	-.01063	SC2B78	.009565
					II4SES	0.00935	SC3IIVEX	.008763	IFACT3	.005272	IE8ADM1	-.00821
					IFACT1	-0.0061	IC1V2	.002916	TA4D4	-.00277	THGI	-0.0023
					SFACT3	.002123	PG1ENROL	.001788	PC9DIST	.001411	PC1USE	8E-06
					IB6RECPT	4E-06						
TA1G13	0.241	TA3BNOCL	-0.4918	1	TA1G46	-.38674	TA1G912	-.17592	TA4D3	.106136	TA1G79	-.15078
		CONST		-	TA4D7	.103665	PA3TLADA	-0.1316	PG1ENROL	-.12919	PA5ISTAT	-.12745
					PB14ATEA	-.11728	TA4D4	.098671	PA6SET	-.09346	SC2B78	-.09286
					IB6AVHDD	-.08698	IC27677	-.07735	TA3ANOST	-.07583	SC2C912	-.06956
					SFACT2	-.06917	TA4D5	0.06903	TA4D1	.068279	PB14BITV	.063558
					PFACT4	-.05558	IC6AVEMK	.054624	IFACT2	.054961	PG2DROPS	-.04316
					SC11LEXP	.043123	PG5ENG2L	-.04081	SC43YRS	-.03748	IFACT1	-.03658
					IE8ADM1	.036277	SC3IIVEX	.035523	IE11IVC0	-.03318	PC3C01R	-.03243
					SD11PRAC	.030545	PFACT1	-.02968	PFACT3	.028403	SH5ENG2L	-.02797
					THGI	-.02745	IF11BRAIN	-.02546	SFACT3	0.02363	IFACT3	.020714
					SB5BITV	.018837	TA4D2	.016956	SB5ATEA	-.01661	TI5ENG2L	-.01505
					SH1EMROL	-.01495	IC1V2	-.01478	PC1USE	.014558	SFACT4	-.01327
					SD7STAFF	-.01322	IC3RGSER	.013086	PG4SES	.012795	SH4SES	.012188
					SA3ADA	-.01093	SFACT1	-.01035	SB3ILIV	-.01004	IFACT4	-.00961
					II1ENROL	-.00789	PFACT2	0.00675	PC9DIST	-.00615	IB6RECPT	.005775
					II4SES	.005126	PB4CILTV	-.00397	SD1RESP	.003799	IC1V1	-.00153
					TA4D4	-7E-04						

Note. Variables in the right-hand dolumns are not included in the regression equation. They are ordered, from left to right, in decreasing order of partial covariances.

Figure 8. PROC IMPUTE Report #4

1. The multiple correlations are estimated from the pairwise correlation coefficients, which are not all based on the same cases, whereas the final parameter estimates are based on only those cases with reported values for the target and all of the predictor variables; multiple correlations calculated from correlations based on different cases should not be interpreted as meaningful, although they prove useful as a tool in accomplishing the imputation;

2. The actual prediction is nonlinear and so may account for more variation than a linear predictor function; and

3. The actual prediction uses discrete levels for the target variable and discrete subsets based on the regression function values, while the multiple R^2 shown in Report #4 is based on a "continuous" predictor function.

The second set of columns shows the predictor variables to be used and the standardized and raw coefficients to be used with each predictor variable. Only variables with significantly nonzero coefficients are included in order to improve cross validation and computational efficiency. The raw regression coefficients give regression function values ranging from zero to the number of regression value subsets selected for this variable. Thus, a simple rounding of the regression value gives the index of the distribution to be used in the final imputation. As a result, the raw regression coefficients do not necessarily yield a value in the same units as the target variable.

The final set of columns show each of the variables not in the equation and a number (labeled PART COV) which, when squared, gives an estimate of the additional percentage of variance (rather than the percentage of additional variance) that would be accounted for if this variable were added to the equation.

In the example TAJANOST (question AJ - how many students does the teacher have) is predicted by PAJTLADA (the total ADA for the school reported by the principal) and by TA1G912 and TA1G79 (whether this is a junior high or high school teacher). The squared multiple correlation is .410; 41% of the variance in TAJANOST is accounted for by these predictors. Of the variables not in the equation, TAJBNOCL (the number of classes taught by this teacher) would improve the prediction the most,

explaining an additional 17% of the variance ($.43^2$). The variable was not included because it had yet to be imputed, so that its value might be missing. In a second equation for TA3ANOST (not shown) , the variable TA3BNOCL was, in fact, included. The next equation shown, in fact, predicts TA3BNOCL from TA3ANOST with a squared multiple R of .726 (the correlation between these two variables is .85).

The final equation in Figure 8 predicts TA1G13, whether the teacher teaches grades 1 through 3. This is predicted by the number of classes taught. The standardized regression coefficient is -.49, meaning that teaching grades 1-3 is predicted by a low number of classes. Since there is a single, discrete predictor, this case is handled a little differently. The regression value subsets will correspond exactly to the distinct values of the predictor variable. As a result the raw regression coefficient has been set to 1 with the constant left undefined.

Missing Data Report #5: Conditional Distributions

Figure 9 shows an example of the fifth report generated by PROC IMPUTE. This report shows the cumulative distribution of each target value in each regression value subset. The first column of this report shows the regression subset number.

The second column gives the number of cases with values for both the target variable and the regression function. This is the number of cases used in estimating the target variable distribution for that subset.

The third column shows the mean regression function value for this subset. This value is used in interpolating between subsets. In the first example predicting TA3BNOCL (number of classes), the first regression value subset included all cases with values below 1.0. The mean regression value of the 699 cases in this subset is .772. The second subset includes cases with regression values between 1.0 and 2.0. These 245 cases have a mean regression value of 1.430. If a case for which TA3BNOCL was missing had a regression function value of 1.101, then the imputed value would be halfway between the value imputed from the subset 1

MISSING DATA REPORT 85: CONDITIONAL DISTRIBUTIONS

REGR. VALUE	N WITH DATA	REGR MEAN	TARGET MEAN	TARGET S.D.	CUMULATIVE PROPORTION FOR EACH TARGET VALUE 1	2	3	4	5	6	7	8
					9	10	11	12	13	14		
TARGET VARIABLE: TA3BNOCL												
1	699	0.772	1.199	0.798	0.921	0.954	0.963	0.978	0.993	1.000	1.000	1.000
					1.000	1.000	1.000	1.000	1.000			
2	245	1.438	2.265	1.565	0.363	0.776	0.824	0.861	0.943	0.976	0.992	1.000
					1.000	1.000	1.000	1.000	1.000			
3	164	2.528	3.744	1.364	0.0	0.220	0.482	0.695	0.884	0.982	0.994	1.000
					1.000	1.000	1.000	1.000	1.000			
4	215	3.497	4.614	1.063	0.0	0.005	0.177	0.423	0.819	0.967	0.995	1.000
					1.000	1.000	1.000	1.000	1.000			
5	265	4.505	5.015	0.695	0.0	0.0	0.019	0.147	0.857	0.974	0.989	1.000
					1.000	1.000	1.000	1.000	1.000			
6	219	5.487	5.242	0.550	0.0	0.0	0.0	0.032	0.753	0.973	1.000	1.000
					1.000	1.000	1.000	1.000	1.000			
7	64	6.494	5.813	0.726	0.0	0.0	0.0	0.031	0.281	0.906	0.969	1.000
					1.000	1.000	1.000	1.000	1.000			
8	18	7.347	6.556	1.002	0.0	0.0	0.0	0.0	0.222	0.667	0.833	0.944
					0.944	0.944	0.944	0.944	1.000			
9	7	8.439	7.000	1.069	0.0	0.0	0.0	0.0	0.0	0.429	0.714	0.857
					1.000	1.000	1.000	1.000	1.000			

TOTAL MEAN= 3.172; S.D.= 2.042; W/I CELL S.D.= 0.998; R SQ= 0.761

TARGET VARIABLE: TA1G13					
1	723	0.0	1.593	0.491	0.407
2	161	1.000	1.354	0.478	0.646
3	104	2.000	1.279	0.448	0.721
4	145	3.000	1.062	0.241	0.938
5	514	4.000	1.025	0.157	0.975
6	195	5.000	1.041	0.198	0.959
7	33	6.000	1.273	0.445	0.727
8	12	7.000	1.417	0.493	0.583
9	1	8.000	1.000	0.0	1.000
10	1	9.000	1.000	0.0	1.000
11	2	10.000	2.000	0.0	0.0
12	0	11.500	0.0	0.0	0.0
13	1	12.000	1.000	0.0	1.000
14	1	13.000	2.000	0.0	0.0

TOTAL MEAN= 1.297; S.D.= 0.457; W/I CELL S.D.= 0.378; R SQ= 0.316

TARGET VARIABLE: PA6SEI						
1	0	0.500	0.0	0.0	0.0	
2	0	1.500	0.0	0.0	0.0	
3	5	2.511	2.600	0.490	0.0	0.400
4	13	3.444	1.615	0.836	0.615	0.769
5	54	4.617	1.444	0.629	0.630	0.926
6	213	5.566	1.432	0.621	0.638	0.930
7	179	6.411	1.637	0.738	0.520	0.844
8	82	7.511	2.134	0.866	0.317	0.549
9	101	8.443	2.505	0.779	0.178	0.317
10	61	9.455	2.607	0.634	0.082	0.311
11	34	10.403	2.412	0.844	0.235	0.353
12	9	11.264	2.667	0.667	0.111	0.222
13	3	12.599	3.000	0.0	0.0	0.0

Cumulative distributions correspond to frequency distributions such as shown in Figure 3.

FIGURE 9. PROC IMPUTE Report #5

distribution and the value imputed from the subset 2 distribution. More generally, if the regression function value is equal to p x (mean for interval i) + (1-p) x (mean for interval i+1), the imputed value would be p x (imputed value from distribution i) + (1-p) x (imputed value from distribution i+1).

The fourth and fifth columns of this report show a mean and standard deviation for the target variable for each regression value subset. For continuous variables, the values shown are the mean and standard deviation of the (integer-valued) level number rather than of the variable itself. In the first example, a discrete variable, the 699 teachers in the first regression value subset taught an average of 1.199 classes while the 7 teachers in the ninth subset taught an average of 7.000 classes.

The remaining columns in this report show the proportion of cases in each subset that have target variable values at or below the indicated level. (The highest level is omitted since all of the cases are at or below this level.) In the example, 92.1% of the teachers in subset 1 taught only one class and all 699 teachers in this subset taught six of fewer classes. The second target variable in the example, teaching grades 1-3, is dichotomous. The number in the rightmost column are the proportions in each subset with a value of zero (the proportion not teaching grades 1-3). Recall from Figure 8 that the single predictor variable is number of classes taught. Here 40.7% of those teaching one class teach other than grades 1-3, while over 90% of those teaching four to six classes teach other than grades 1-3.

The row at the bottom of each table in this report shows the overall mean and standard deviation of the target variable (in integer level units) and the average standard deviation within each subset. The R SQ measure (actually an eta squared since the relationship may not be linear) indicates the reduction in variance due to the differences between subsets.

Mean Square Error Due to Imputation

The final report generated by PROC IMPUTE is not yet fully developed. It is designed to show the error variation in each variable due to missing values. Each time a variable is imputed, the target variable variance for the appropriate regression value subset is added to a total for that variable. There is no missing data error variation for nonmissing values so nothing is added to the totals for these cases. The final totals are then divided by the total number of cases to give an average missing data error variance for each variable. If no cases were missing values, then the average will be zero. Similarly if all missing values are imputed with certainty (the within subset variances were all zero), then the final average error would be zero.

The final error variance estimates are printed at the end of Report #5. A number is given for each variable in the VAR list and the numbers are in order of the variable's position in this list. The variances shown are currently in integer level units and must be referenced to the Total S.D. in Report #5. This measure can be used to assess the random component of the error due to imputing a value rather than collecting real data. An R SQ less than .25, for a target variable with substantial missing data and for which nonrespondents differ significantly from respondents (Report #2), indicates generally poor imputation of the target variable.

Processing Time by Numbers of Variables and Cases for BMDPAM and Proc IMPUTE*

	Number of Variables						
	10		20		30	46	67
Number of Cases	BMDPAM	IMPUTE	BMDPAM	IMPUTE	IMPUTE	IMPUTE	IMPUTE
100		.9					
500	3.2	2.0	7.0	4.6	8.3		
1000			11.8	8.2			
1179						37.5	
1994							105.2

*Processing in time is in CPU seconds for an IBM 370/168 in an MVS environment. The BMDPAM runs used the REGR option.

APPENDIX B

Sample SAS Program to Reweight for Total Nonresponse

```
STEP
1       DATA TEMP1;
            SET DDNAME1.YOUR FILE;
            * INCLUDE ANY CØDE NEEDED TØ SET THE          *;
            * 'CELL', 'NØNRESP', AND 'WEIGHT' VARIABLES*;
            * ALSO TO CØMBINE CELL AS NEEDED             *;
            * NOTE: CELL = UNIQUE FOR EACH CELL          *;
            *       NØNRESP=1 FØR NØNRESPØNDENTS,
            *              0 ØTHERWISE                   *;
            *       WEIGHT =CASE WEIGHT TO BE RESET      *;
2           PRØC SØRT; BY CELL NØNRESP;
3           PRØC MEANS; BY CELL NØNRESP;
            VAR WEIGHT;
            ØUTPUT  ØUT = TEMP2
               SUM = SUMWT;
4           DATA WTADJS; BY CELL;
            RETAIN SUMRESP 0;
            IF LAST.CELL THEN GØ TØ SETWT;
            SUMRESP = SUMWT; DELETE;
            SETWT:
            IF SUMWT LE 0 THEN GØ TØ CELLERR;
            WTADJ = (SUMMWT + SUMRESP) / SUMRESP;
            SUMRESP = 0;
            KEEP CELL WTADJ; RETURN;
            CELLERR:
            PUT CELL= 'HAS NØ RESPØNDENTS BUT HAS'
                SUMWT 'WEIGHTED NØNRESPØNDENTS';
5           PRØC PRINT;  * TØ PRINT WEIGHT ADJUSTMENTS;
6           DATA DDNAME2.NEWFILE;
            MERGE TEMP1 WTADJS; BY CELL;
            IF NØNRESP EQ 1 THEN DELETE;
            WEIGHT = WEIGHT * WTADJ;
```

REFERENCES

Aziz, F., & Scheuren, F. (Eds.). Imputation and editing of faulty or missing survey data. U.S. Department of Commerce, 1978.

Cox, B., & Folsom, R. An empirical investigation of alternative item nonresponse adjustment procedures. National Center for Education Statistics, 1979.

Dixon, W., & Brown, M. (Eds.). BMDP-77 biomedical computer programs: P-series. Los Angeles: University of California Press, 1977.

Helwig, J., & Council, K. (Eds.). SAS user's guide (1979 edition). Raleigh, N.C.: SAS Institute, Inc., 1979.

Madow, W. (Ed.). Symposium on incomplete data: Preliminary proceedings. U.S. Social Security Administration, 1979.

Nie, N.; Hull, C.; Jenkins, J.; Steinbrenner, K.; & Bent, D. Statistical package for the social sciences (second edition). New York: McGraw-Hill, 1975.

Oh, H., & Scheuren, F. Multivariate raking ratio estimation in the 1973 Exact Match study. In F. Aziz & F. Scheuren (Eds.), Imputation and editing of faulty or missing survey data. U.S. Department of Commerce, 1978.

Rubin, D. Multiple imputations in sample surveys--a phenomenological Bayesian approach to nonresponse. In F. Aziz & F. Scheuren (Eds.), Imputation and editing of faulty or missing survey data. U.S. Department of Commerce, 1978.